METHOD AND DEVICE FOR ENCODING WIDEBAND SPEECH, ALLOWING IN PARTICULAR AN IMPROVEMENT IN THE QUALITY OF THE VOICED SPEECH FRAMES

Field of the Invention

The present invention relates to the encoding/decoding of wideband speech, and in particular, with respect to mobile telephony.

Background of the Invention

In wideband speech, the bandwidth of the speech signal lies between 50 and 7,000 Hz. Successive speech sequences sampled at a predetermined sampling frequency, for example 16 kHz, are processed in a coding device of the CELP type using coded-sequence-excited linear prediction. For example, one such device is referred to as ACELP, which stands for algebraic code excited linear prediction. This device is well known to one skilled in the art, and is described in recommendation ITU-TG 729, version 3/96, entitled "Coding Of Speech At 8 kbits/s By Conjugate Structure-Algebraic Coded Sequence Excited Linear Prediction".

The main characteristics and functions of such a coder will now be briefly discussed while referring to Figure 1. Further details may be found in the above mentioned recommendation.

The prediction coder CD of the CELP type is based on the model of code-excited linear predictive

coding. The coder operates on voice super-frames equivalent to 20 ms of signal for example, and each comprises 320 samples. The extraction of the linear prediction parameters, that is, the coefficients of the linear prediction filter which is also referred to as the short-term synthesis filter 1/A(z), is performed for each speech super-frame. Each super-frame is subdivided into frames of 5 ms comprising 80 samples. For every frame, the voice signal is analyzed to extract therefrom the parameters of the CELP prediction model.

In particular, the extracted parameters include a long-term excitation digital word v_i extracted from an adaptive coded directory also referred to as an adaptive long-term dictionary LTD, an associated long-term gain Ga, a short-term excitation word c_j extracted from a fixed coded directory also referred to as a short-term dictionary STD, and an associated short-term gain Gc.

These parameters are thereafter coded and transmitted. At reception, these parameters are used in a decoder to recover the excitation parameters and the predictive filter parameters. The speech is then reconstructed by filtering the excitation stream in a short-term synthesis filter.

The adaptive dictionary LTD contains digital words representative of tonal lags representative of past excitations. The short-term dictionary STD is based on a fixed structure, for example of the stochastic type or of the algebraic type, using a model involving an interleaved permutation of Dirac pulses. In the case of an algebraic structure, the coded directory contains innovative excitations also referred

to as algebraic or short-term excitations. Each vector contains a certain number of non-zero pulses, for example four, each of which may have the amplitude +1 or -1 with predetermined positions.

The processing means of the coder CD functionally comprises first extraction means MEXT 1 for extracting the long-term excitation word, and second extraction means MEXT 2 for extracting the short-term excitation word. Functionally, the extraction means MEXT 1 and MEXT 2 are embodied in software within a processor for example.

The extraction means MEXT 1 and MEXT 2 each comprise a predictive filter PF having a transfer function equal to $1/A(z)$, as well as a perceptual weighting filter PWF having a transfer function $W(z)$. The perceptual weighting filter PWF is applied to the signal to model the perception of the ear. Furthermore, the extraction means MEXT 1 and MEXT 2 each comprise means MSEM for performing a minimization (i.e., a reduction) of a mean square error.

The synthesis filter PF of the linear prediction models the spectral envelope of the signal. The linear prediction analysis is performed every super-frame to determine the linear predictive filtering coefficients. The latter are converted into pairs of spectral lines, i.e., line spectrum pairs LSP, and are digitized by predictive vector quantization in two steps.

Each 20 ms a speech super-frame is divided into four frames of 5 ms each containing 80 samples. The quantized LSP parameters are transmitted to the decoder once per super-frame, whereas the long-term and short-term parameters are transmitted at each frame.

The quantized and non-quantized coefficients of the linear prediction filter are used for the most recent frame of a super-frame, while the other three frames of the same super-frame use an interpolation of these coefficients. The open-loop tonal lag is estimated, for example every two frames on the basis of the perceptually weighted voice signal. The following operations are repeated at each frame.

The long-term target signal X_{LT} is calculated by filtering the sampled speech signal s(n) by the perceptual weighting filter PWF. The zero-input response of the weighted synthesis filters PF and PWF is thereafter subtracted from the weighted voice signal to obtain a new long-term target signal. The impulse response of the weighted synthesis filter is calculated.

A closed-loop tonal analysis using minimization or reduction of the mean square error is thereafter performed to determine the long-term excitation word v_i and the associated gain Ga by the target signal and of the impulse response, and by searching around the value of the open-loop tonal lag.

The long-term target signal is thereafter updated by subtraction of the filtered contribution y of the adaptive coded directory LTD. This new short-term target signal X_{ST} is used during the exploration of the fixed coded directory STD to determine the short-term excitation word c_j and the associated gain G_c. Here again, this closed-loop search is performed by minimization of the mean square error.

The adaptive long-term dictionary LTD as well as the memories of the filters PF and PWF are updated by the long-term and short-term excitation words thus

determined. The quality of a CELP algorithm depends strongly on the richness of the short-term excitation dictionary STD, for example an algebraic excitation dictionary. Even though the effectiveness of such an algorithm is very high for narrow bandwidth signals (300-3,400 Hz), problems arise with respect to wideband signals.

Even with a very rich dictionary, the speech encoding algorithm produces a reconstructed signal corrupted by various types noise, and in particular, a whistling type noise that mars voiced speech frames. This high-frequency noise stems from the short-term excitation that introduces undesirable artifacts. Two types of approaches for addressing this problem have already been proposed.

A first approach proposes that the short-term contribution be rendered periodic. This is described for example in the following articles by Gerson and Jasiuk, entitled "Techniques For Improving The Performance Of CELP-Type Speech Coders", IEEE, Journal on Selected Areas in Communications, Vol. 10, No 5, June 1992, pages 858-865; and by Miki et al., entitled "A Pitch Synchronous Innovation CELP (PSI-CELP) Coder For 2-4 kbit/s", Proc., IEEE Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP'84, Adelaide, South Australia, 1994, Vol. II, pages 113-116.

The other approach proposes that the short-term gain be adaptively controlled. This is described for example, in the following articles by Taniguchi, Johnson and Ohta, entitled "Pitch Sharpening For Perceptually Improved CELP, And The Sparse-Delta Codebook For Reduced Computation", Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,

ICASSP'91, Toronto, Canada, 1991, pages 241-244; and by Shoham, entitled "Constrained-Stochastic Excitation Coding Of Speech At 4.8 kbit/s", Advances in Speech Coding, B.S. Atal, V. Cuperman, and A. Gersho, Eds., Dordrecht, The Netherlands, Kluwer, 1991, pages 339-348.

Summary of the Invention

In view of the foregoing background, an object of the present invention is to provide a wideband speech encoding method in which the speech is sampled for obtaining successive voice frames each comprising a predetermined number of samples, and parameters of a code-excited linear prediction model are determined for each voice frame. These parameters may include a long-term excitation digital word extracted from an adaptive coded directory and an associated long-term gain, and include a short-term excitation word extracted from a short-term dictionary and an associated short-term gain. The adaptive coded directory may be updated based upon the extracted long-term excitation word and the extracted short-term excitation word.

According to a general characteristic of the invention, the method comprises an updating of the state of the linear prediction filter with the short-term excitation word filtered by a filter of an order greater than or equal to 1. For example, the filter may be a finite impulse response filter of order 1 whose coefficients depend on the value of the long-term gain to reduce the contribution of the short-term excitation when the gain of the long-term excitation is

greater than a predetermined threshold, such as 0.8 for example.

The method according to the present invention includes weakening or reducing the contribution of the short-term excitation if the gain of the long-term excitation is large. However, it is the contribution of the unweakened short-term excitation that is stored in the adaptive dictionary for its updating. Thus, the reduction occurs only on the output. It is important to preserve the short-term contribution to be stored, since the richness of the adaptive dictionary is thus maintained for the lowest frequencies. Of course, weakening of the contribution may also be applied during the reconstruction of the signal at the decoder level.

According to one mode of implementation, in which the filter is of an order 1 and has a transfer function equal to $B0+B1\ z^{-1}$, the first coefficient B0 of the filter is equal to $1/(1+\beta.\min(Ga,1))$, and the second coefficient B1 of the filter is equal to $\beta.\min(Ga,1)/(1+\beta.\min(Ga,1))$, where β is a real number of absolute value less than 1, Ga is the long-term gain and min(Ga,1) designates the minimum value between Ga and 1.

According to one variation of the invention, the extraction of the long-term excitation word is performed using a first perceptual weighting filter comprising a first formantic weighting filter, and the extraction of the short-term excitation word is performed using the first perceptual weighting filter cascaded with a second perceptual weighting filter. The second perceptual weighting filter comprises a second formantic weighting filter. The denominator of the

transfer function of the first formantic weighting filter is equal to the numerator of the second formantic weighting filter.

Thus, according to this variation, the use of two different formantic weighting filters makes it possible to control the short-term and the long-term distortions independently. The short-term weighting filter is cascaded with the long-term weighting filter. The present invention thus provides an approach similar to the gain control type as discussed above, but is totally different from that described in the articles by Taniguchi et al. and by Shoham.

Tying of the denominator of the long-term weighting filter to the numerator of the short-term weighting filter makes it possible to control these two filters separately, and allows a significant simplification when these two filters are cascaded. There may also be a provision for an updating of the state of the two perceptual weighting filters with the short-term excitation word filtered by the filter having an order greater than or equal to 1.

Another aspect of the present invention is directed to a wideband speech encoding device comprising sampling means for sampling the speech in such a way as to obtain successive voice frames each comprising a predetermined number of samples. The device may further comprise processing means for determining parameters of a code-excited linear prediction model for each voice frame. The processing means may comprise first extraction means for extracting a long-term excitation digital word from an adaptive coded directory and for calculating an associated long-term gain, and second extraction means

for extracting a short-term excitation word from a fixed coded directory and for calculating an associated short-term gain. The device may comprise first updating means for updating the adaptive coded directory on the basis of the extracted long-term excitation word and of the extracted short-term excitation word.

The first extraction means may comprise a linear prediction digital filter, and the device may comprises second updating means for updating the state of the linear prediction filter with the short-term excitation word filtered by a filter. The filer has an order greater than or equal to 1 whose coefficients depend on the value of the long-term gain to weaken the contribution of the short-term excitation when the gain of the long-term excitation is greater than a predetermined threshold.

The first extraction means may further comprise a first perceptual weighting filter and a first formantic weighting filter, and the second extraction means may comprise the first perceptual weighting filter cascaded with a second perceptual weighting filter which comprises a second formantic weighting filter. The denominator of the transfer function of the first formantic weighting filter may be equal to the numerator of the second formantic weighting filter.

Yet another aspect of the present invention is directed to a terminal of a wireless communication system, for example a cellular mobile telephone incorporating a device as defined above.

Brief Description of the Drawings

Other advantages and characteristics of the invention will become apparent on examining the detailed description of embodiments and modes of implementation, which are in no way limiting, and the appended drawings, in which:

Figure 1 diagrammatically illustrates a speech encoding device according to the prior art;

Figures 2 and 2a diagrammatically illustrate an encoding device and a corresponding decoder according to the present invention;

Figure 3 diagrammatically illustrates another embodiment of an encoding device according to the present invention; and

Figure 4 diagrammatically illustrates the internal architecture of a cellular mobile telephone incorporating a coding device according to the present invention.

Detailed Description of the Preferred Embodiments

The encoding device or coder CD according to the invention, as illustrated in Figure 2, is distinguished from that of the prior art as illustrated in Figure 1 by the fact that the coder CD further comprises second updating means UPD2 for performing an updating of the state of the linear prediction filter PF, and of the state of the perceptual weighting filter PWF with the short-term excitation word c_j filtered by a filter FLT1 having an order greater than or equal to 1. This filter may be a finite impulse response filter of order 1, for example.

The coefficients of this filter of order 1 depend on the value of the long-term gain Ga to weaken

the contribution of the short-term excitation when the gain of the long-term excitation Ga is greater than a predetermined threshold, such as equal to 0.8, for example.

By way of example, the transfer function of the filter FLT1 is equal to $B0+B1\ z^{-1}$ and the first coefficient of the filter B0 may be determined through the formula (I) below:

$$1/(1 + 0.98 \min (Ga, 1)) \qquad (I)$$

whereas the second coefficient of the filter B1 may be determined through the formula (II) below:

$$0.98 \min (Ga, 1)/(1 + 0.98 \min (Ga, 1)) \qquad (II)$$

It is actually the unweakened short-term contribution which is stored in the adaptive dictionary LTD for its updating.

Thus, the weakening intervenes only on the output signal, and by retaining the contribution of the short-term excitation to be stored it is possible to preserve the richness of the adaptive dictionary for the lowest frequencies.

Naturally, the filtering of the excitation must also be applied with respect to the updating of the state of the memories of the filters in the decoder DCD, as illustrated diagrammatically in Figure 2a. The embodiment illustrated in Figure 2 makes it possible to eliminate a whistling type noise in the voiced speech frames.

The perceptual weighting filter PWF utilizes the masking properties of the human ear with respect to the spectral envelope of the speech signal, the shape

of which depends on the resonances of the vocal tract. This filter makes it possible to attribute more importance to the error appearing in the spectral valleys as compared with the formantic peaks.

In the variation illustrated in Figure 2, the same perceptual weighting filter PWF is used for the short-term and long-term search. The transfer function W(z) of this filter PWF is given by the formula (III) below:

$$W(z) = \frac{A(z / \gamma_1)}{A(z / \gamma_2)} \qquad \text{(III)}$$

in which 1/A(z) is the transfer function of the predictive filter PF and $\gamma 1$ and $\gamma 2$ are the perceptual weighting coefficients, the two coefficients being positive or zero and less than or equal to 1 with the coefficient $\gamma 2$ less than or equal to the coefficient $\gamma 1$.

In a general manner, the perceptual weighting filter is constructed from a formantic weighting filter and from a filter for weighting the slope of the spectral envelope of the signal (tilt). In the present case, it will be assumed that the perceptual weighting filter is formed only from the formantic weighting filter whose transfer function is given by formula (III) above.

The spectral nature of the long-term contribution is different from that of the short-term contribution. Consequently, it is advantageous to use two different formantic weighting filters, making it possible to control the short-term and long-term distortions independently.

Such an embodiment is illustrated in Figure 3, in which, as compared with Figure 2, the single filter PWF has been replaced by a first formantic weighting filter PWF1 for the long-term search, cascaded with a second formantic weighting filter PWF2 for the short-term search. Since the short-term weighting filter PWF2 is cascaded with the long-term weighting filter, the filters appearing in the long-term search loop must also appear in the short-term search loop.

The transfer function $W_1(z)$ of the formantic weighting filter PWF1 is given by formula (IV) below:

$$W_1(z)=\frac{A(z/\gamma_{11})}{A(z/\gamma_{12})} \qquad \text{(IV)}$$

whereas the transfer function $W_2(z)$ of the formantic weighting filter PWF2 is given by formula (V) below:

$$W_2(z)=\frac{A(z/\gamma_{21})}{A(z/\gamma_{22})} \qquad \text{(V)}$$

The coefficient γ_{12} is equal to the coefficient γ_{21}. This allows a significant simplification when these two filters are cascaded. Thus, the filter equivalent to the cascade of these two filters has a transfer function given by the formula (VI) below:

$$\frac{A(z/\gamma_{11})}{A(z/\gamma_{22})} \qquad \text{(VI)}$$

If one uses the value 1 for the coefficient γ_{11}, then the synthesis filter PF having the transfer function $1/A(z)$ followed by the long-term weighting filter PWF1

and by the weighting filter PWF2, it is then equivalent to the filter whose transfer function is given by the formula (VII) below:

$$\frac{1}{A(z / \gamma_{22})} \qquad \text{(VII)}$$

This further considerably reduces the complexity of the algorithm for extracting the excitations. For example, it is possible to use the respective values 1, 0.1 and 0.9 for the coefficients γ_{11}, $\gamma_{21} = \gamma_{12}$ and γ_{22}.

The invention applies advantageously to mobile telephony, and in particular to any remote terminal belonging to a wireless communication system. Such a terminal, for example a mobile telephone TP as illustrated in Figure 4, conventionally comprises an antenna linked by a duplexer DUP to a reception chain CHR and to a transmission chain CHT. A baseband processor BB is linked respectively to the reception chain CHR and to the transmission chain CHT by an analog-to-digital converter ADC and a digital-to-analog converter DAC.

Conventionally, the processor BB performs baseband processing, and in particular a channel decoding DCN, followed by a source decoding DCS. For transmission, the processor performs a source coding CCS followed by a channel coding CCN. When the mobile telephone incorporates a coder according to the invention, the latter is incorporated within the source coding means CCS, whereas the decoder is incorporated within the source decoding means DCS.